Neural Computing and Applications (2020) 32:10209–10228
https://doi.org/10.1007/s00521-019-04559-1

ORIGINAL ARTICLE

Gated multimodal networks

John Arevalo[1] · Thamar Solorio[2] · Manuel Montes-y-Gómez[3] · Fabio A. González[1]

Received: 8 May 2019 / Accepted: 5 October 2019 / Published online: 15 January 2020

Abstract
This paper considers the problem of leveraging multiple sources of information or data modalities (e.g., images and text) in neural networks. We define a novel model called gated multimodal unit (GMU), designed as an internal unit in a neural network architecture whose purpose is to find an intermediate representation based on a combination of data from different modalities. The GMU learns to decide how modalities influence the activation of the unit using multiplicative gates. The GMU can be used as a building block for different kinds of neural networks and can be seen as a form of intermediate fusion. The model was evaluated on two multimodal learning tasks in conjunction with fully connected and convolutional neural networks. We compare the GMU with other early- and late-fusion methods, outperforming classification scores in two benchmark datasets: MM-IMDb and DeepScene.

Keywords Multimodal learning · Representation learning · Information fusion · GMU

1 Introduction

Representation learning methods have received a lot of attention by researchers and practitioners because of their successful application to complex problems in areas such as computer vision, speech recognition and text processing [40]. Most of these efforts have concentrated on data involving one type of information (images, text, speech, etc.), despite data being naturally multimodal. Multimodality refers to the fact that the same real-world concept can be described by different views or data types. Collaborative encyclopedias (such as Wikipedia) describe a famous person through a mixture of text, images and, in some cases, audio. Users from social networks comment about events like concerts or sport games with small phrases and multimedia attachments (images/videos/audios). Patient's medical records are represented by a collection of images, text, sound and other signals. The increasing availability of multimodal databases from different sources has motivated the development of automatic analysis techniques to exploit the potential of these data as a source of knowledge in the form of patterns and structures that reveal complex relationships [7, 11]. In recent years, multimodal tasks have received attention by the representation learning community. Strategies for visual question answering [5] or image captioning [31, 67, 72] have developed interesting ways of combining different representation learning architectures.

Most of these models are focused on mapping from one modality to another or solving an auxiliary task to create a common representation with the information of all modalities. In this work, we design a novel module that combines multiple sources of information, which is optimized with respect to the end goal objective function. Our proposed module is based on the idea of using gates for combining input modalities giving a higher importance to the ones that are more likely to contribute for correctly

✉ John Arevalo
jearevaloo@unal.edu.co

Thamar Solorio
solorio@cs.uh.edu

Manuel Montes-y-Gómez
mmontesg@inaoep.mx

Fabio A. González
fagonzalezo@unal.edu.co

[1] Department of Computing Systems and Industrial Engineering, Universidad Nacional de Colombia, Cra 30 No 45 03-Ciudad Universitaria, Bogotá, Colombia

[2] Department of Computer Science, University of Houston, Houston, TX 77204-3010, USA

[3] Computer Science Department, Instituto Nacional de Astrofísica, Óptica y Electrónica, C.P. 72840 Puebla, Mexico

generating the desired output. We use multiplicative gates that assign importance to various features simultaneously, creating a rich multimodal representation that does not require manual tuning, but instead it learns directly from the training data. We show in the experimental evaluation that our gated model can be reused in different network architectures for solving different tasks, and can be optimized end-to-end with other modules in the architecture using standard gradient-based optimization algorithms. Such behavior was evidenced in the experimental analysis that suggested that the gain is based on giving more weight to specific modalities for specific problems.

We explored two application use cases: genre movie prediction, and image segmentation. On the one hand, genre prediction has several application areas like document categorization [32], recommendation systems [47], and information retrieval systems, among others. On the other hand, image segmentation is heavily used in autonomous drive systems [62], medical imaging [30] and other computer vision tasks. The motivation to chose the above tasks is twofold: (1) to evaluate the model in different and unrelated scenarios in order to support that the model is suitable for different multimodal learning tasks, and (2) to integrate the proposed unit in the most popular network architectures: convolutional and fully connected.

The main contribution of this work is a new deep neural network building block, the gated multimodal unit (GMU), which is able to learn an input-dependent gate-activation pattern that determines how each modality contributes to the output of hidden units. This generalizes conventional multimodal late- and early-fusion architectures to a modular intermediate fusion that can be used in different stages of a neural network, combined with other layer types (e.g., convolutional or recurrent), and trained in an end-to-end fashion.

The rest of this paper is organized as follows: Sect. 2 gives an overview of previous related work. Section 3 presents the gated multimodal unit (GMU) and empirically evaluates its behavior with synthetic experiments. Section 4 presents a systematic evaluation in two supervised learning tasks. Finally, Sect. 5 summarizes the main aspects of this work and presents its main conclusions.

2 Related work

Different reviews [7, 11, 17, 41] have summarized strategies that addressed multimodal analysis. Most of the reviewed works claimed the superiority of multimodal over unimodal approaches for automatic analysis tasks. A conventional multimodal analysis system receives as input two or more modalities that describe a particular object. The most common multimodal sources are video, audio, images and text. In recent years there has been a consensus with respect to the use of representation learning models to characterize the information of this kind of sources [40]. However, the way that such extracted features are combined is still in exploration.

Multimodal combination seeks to generate a single representation that eases automatic analysis tasks when building classifiers or other predictors. A basic approach is to concatenate features to get a final representation [34, 55, 61]. Although it is a straightforward strategy, given that the nature of data for each modality is different, their statistical properties usually are not shared across modalities [59], and thus the predictor needs to model complex interactions between them. Instead, more elaborated combination strategies have been proposed, in which prior knowledge is exploited, additional information is included or multimodal interactions are explicitly modeled. Some of those strategies include Restricted Boltzmann Machines (RBMs) and autoencoders [53] was one of the first multimodal methods based on deep architectures. The model concatenated higher level representations and trained two RBMs to reconstruct the original audio and video representations, respectively. Additionally, they trained a model to reconstruct both modalities given only one of them as input. In an interesting result, Ngiam et al. [53] were able to mimic a perceptual phenomenon that demonstrates an interaction between hearing and vision in speech perception known as McGurk effect. However, notice that RBMs limits its scalability because a Monte Carlo Markov chain is required during training stage. A similar approach was proposed by Srivastava and Salakhutdinov [59]. They modified feature learning and reconstruction phases with Deep Boltzmann Machines. Authors claimed that such strategy is able to exploit unlabeled data by improving the performance in retrieval and annotation tasks. Other similar strategies propose to fuse modalities using neural network architectures [4, 19, 33, 37, 45, 50, 63, 70] with two input layers separately and including a final supervised layer such as softmax regression classifier.

An alternative approach involves an objective or loss function suited for the target task [1, 21, 38, 50, 57, 58, 78]. These strategies usually assume that there exists a common latent space where modalities can express the same semantic concept through a set of transformations of the raw data. The semantic embedding representations are such that two concepts are similar if and only if their semantic embeddings are close [54]. In [57] a multimodal strategy to perform zero-shot classification was proposed. They trained a word-based neural network model [24] to represent textual information, while use unsupervised feature learning models proposed in [16] to get image representation. The fusion was done by learning an image linear mapping to project images into the semantic word space

learned in the neural network model. Additionally a Bayesian framework was included to decide whether an image is of a seen or unseen class. Frome et al. [21] learn the image representation using a CNN trained with the Imagenet dataset and a word-based neural language model [52] to represent the textual modality. To perform the fusion they re-train the CNN using the text representation as target. This work outperforms the scalability of [57] from 2 to 20,000 unknown classes in the zero-shot learning task. A modified strategy of Frome et al. [21] was presented by Norouzi et al. [54]. Instead of re-training the CNN network, they built a convex combination with probabilities estimated by the classifier and semantic embedding vector of the unseen label. This simple strategy outperforms state-of-the-art results. Because the cost function involves both multimodal combination and supervision, these family of models are tied to the task of interest. Thus, if the domain or task conditions change, an adaptation of the model is required.

Bayesian alternatives to combine information also have been explored. In particular, Bayesian fusion has been applied to multispectral images as proposed in [68, 69]. Despite their interesting results, it should be noticed that this family of methods require a Monte Carlo Markov chain process to train the model, making it harder to scale in comparison with gradient-based methods. This approach also requires additional adaptation when is applied in a different task.

The proposed model is closely related to the mixture of experts (MoE) approach [29]. However, the common usage of MoE is focused on performing decision fusion, i.e., combining predictors to address a supervised learning problem [76]. Similar late-fusion models have been extended to deep architectures with bagging methods [2]. Our model is devised as a new component in the representation learning scheme, making it independent from the final task (e.g., classification, regression, unsupervised learning, etc) provided that the defined cost function be differentiable. On the other hand, It is noteworthy that extending current models to deal with more than two modalities is a complex challenge [77]. Our proposed method addressed this multimodal challenge by generalizing the gate approach with independent parameters per modality.

GMUs were presented for the first time at [6] as a working paper that was not formally published. This paper extends such work by performing a more systematic experimental evaluation and introducing a new use case in a computer vision task where GMUs are integrated in a convolutional architecture.

Movie genre prediction is a multilabel task since most of the movies belong to more than one genre, (e.g., Matrix (2000) is a Sci-fi/Action movie). In this setup, Anand [3] explores the efficiency of using keywords and users' tags to perform multilabeling using the movies from MovieLens 1M dataset which contains 1700 movies. Also Ivasic-Kos et al. [27, 28] performed multilabel classification using handcrafted features from posters, with 1500 samples for six genres. Makita and Lenskiy [47, 48] use movie ratings matrix and genre correlation matrix to predict the genre. It used a smaller version of the Movielens dataset with 18 movie genres. Most of the above works have used the publicly available MovieLens datasets. However, there is not a single experimental setup defined so that all methods can be systematically compared. Also, to the best of our knowledge, none of the previous works contain more than 10, 000 samples. With this work we released a dataset created with the movies of the MovieLens 20M dataset. We include not only genre, poster and plot information used in this work, but also the poster of the movie as well as more than 50 characteristics taken from the IMDb website.

Multimodal image segmentation has been addressed with representation learning techniques using RGB and depth images. Pei et al. [55] learned a dictionary from concatenated patches from RGB and depth images to extract features from small regions, then those features are used to train a pixel-based classifier. In a similar setup, Valada et al. [64] integrated a mixture of experts model in a convolutional neural network to segment six concepts in outdoor images. They explored different modalities, obtaining their best results when RGB and depth images were combined. Our work is similar because it is also an end-to-end convolutional neural network, trained with gradient-based algorithms, but differs in the way the modalities are fused. While [64] used two predictors to combine the information, we instead used gates to combine intermediate representations. This allows our model to be applied also in unsupervised tasks such as image generation or feature learning, provided that the model can be trained with gradient-based approaches.

3 Methods

This paper presents a neural-network-based strategy for addressing supervised tasks with multimodal data. The key component of such strategy is a novel type of hidden unit, the Gated Multimodal Unit (GMU), which learns to decide how modalities influence the activation of the unit using gates. The first part of this section exposes the details of the GMU, while the second part analyzes its behavior in a synthetic scenario.

3.1 Gated multimodal unit

Multimodal learning is closely related to data fusion. Data fusion looks for optimal ways of combining different information sources into an integrated representation that provides more information than the individual sources [11]. This fusion can be performed at different levels and can be categorized into two broad categories: feature fusion and decision fusion. Feature fusion, also called early fusion, looks for a subset of features from different modalities, or combinations of them, that better represent the information needed to solve a particular problem. On the other hand, decision fusion, or late fusion, combines decisions from different systems, e.g., classifiers, to produce consensus. This consensus may be reached by a simple average, a voting system or a more complex Bayesian framework.

In this work we present a model, based on gated neural networks, for data fusion that combines ideas from both feature and decision fusion. The model, called Gated Multimodal Unit (GMU), is inspired by the control flow in recurrent architectures like gated recurrent units [14] or the long short-term memory unit [23]. A GMU is intended to be used as an internal unit in a neural network architecture whose purpose is to find an intermediate representation based on a combination of data from different modalities. Figure 1a depicts the structure of a GMU. Each x_i corresponds to a feature vector associated with modality i. Each feature vector feeds a neuron with a tanh activation function, which is intended to encode an internal representation feature based on the particular modality. For each input modality, x_i, there is a gate neuron (represented by σ nodes in the diagram), which controls the contribution of the feature calculated from x_i to the overall output of the unit. When a new sample is fed to the network, a gate neuron associated to modality i receives as input the feature vectors from all the modalities and uses them to decide the degree of contribution of the modality i to the internal encoding of the particular input sample.

Figure 1b shows a simplified version of the GMU for two input modalities, x_v (visual modality) and x_t (textual modality). It should be noted that models from Fig. 1a, b are not completely equivalent, since in the bimodal case the gates are tied. Such weight tying constraints the model, so that the units control the trade-off between both modalities while they use less parameters than the multimodal case. For sake of clarity, below we detailed the equations governing a single GMU. Notice that in practice is common to have multiple units in the same layer.

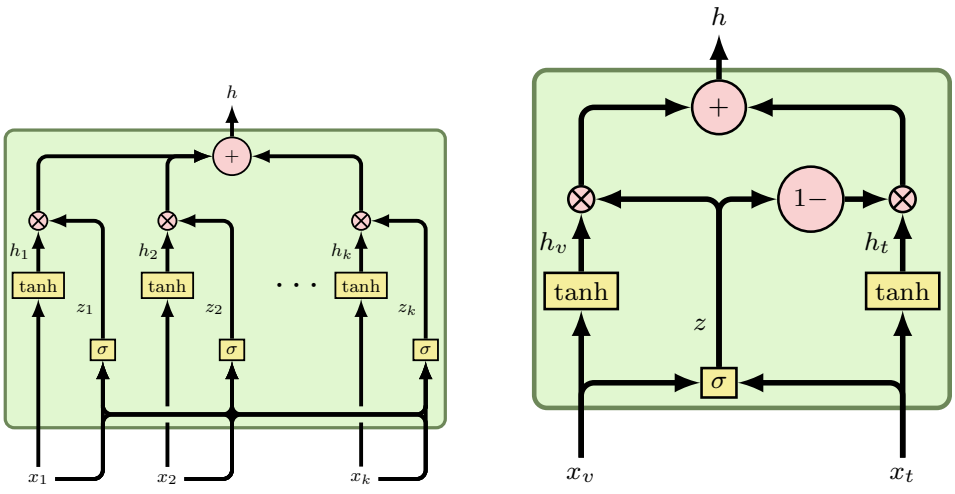
Let $x_v \in \mathbb{R}^{d_v}$, $x_t \in \mathbb{R}^{d_t}$ be the column vectors representing the visual and textual modalities, respectively. The GMU extracts hidden features for each modality as follows:

$$h_v = \tanh\left(W_v x_v^\top\right)$$
$$h_t = \tanh\left(W_t x_t^\top\right)$$

where $W_v \in \mathbb{R}^{d_v}$ and $W_t \in \mathbb{R}^{d_t}$ are the learnable weights, tanh is the default activation function and, $h_t, h_v \in \mathbb{R}$ are the resultant hidden representations. The GMU contains a third internal feature $z \in \mathbb{R}$, calculated as follows:

$$z = \sigma\left(W_z[x_v, x_t]^\top\right)$$

where $[\cdot,\cdot]$ denotes the concatenation operator, $W_z \in \mathbb{R}^{d_v+d_t}$ are the learnable weights and σ represents the sigmoid activation function. The output activation, $h \in \mathbb{R}$, of the GMU is given by a convex combination of h_t and h_v weighted by the z activation:



(a) The model to use with more than two modalities.

(b) A simplification for the bimodal approach.

Fig. 1 Illustration of a single gated unit for bimodal and > 2 modalities. $x, z, h \in \mathbb{R}$, yellow boxes represent the activation functions $f : \mathbb{R} \to \mathbb{R}$. Red circles with cross sign represent element-wise multiplication, red circles with $+$ sign represent summation of all the inputs and red circle with "$1-$" represents the function $f(s) = 1 - s$ (colour figure online)

$$h = zh_v + (1 - z)h_t$$

This formulation allows the GMU to decide how each modality affects the unit's output. This also means that each different input will have different weights in such convex combination due to the dependency of z on x_v and x_t. Since all are differentiable operations, this model can be easily coupled with other neural network architectures and trained with stochastic gradient descent.

3.2 Noisy channel model

In order to analyze the behavior of the GMU, we built a synthetic scenario to determine which modality carries the most relevant information. Consider the channel model illustrated in Fig. 2. There is an original source signal that is transformed by two independent components T_1 and T_2. The signals from T_1 and T_2 are transmitted by two channels, C_1 and C_2, respectively, that have two operation modes. In mode one, the channel transmits the original signal, in mode two, it transmits noise. A switch controls which channel will carry the signal. In one position, C_1 carries the signal and C_2 carries noise, in the other position, the situation is inverted. The switch may change its position at any time. The goal is to get the information of the original signal from the combination of the signals C_1 and C_2 without knowing which one is carrying the information and which one is carrying noise at a given time.

We implemented the noisy channel scenario through the generative model depicted in Fig. 3. In this model we define the random binary variable C as the target and $x_v, x_t \in \mathbb{R}$ as the input features. M is a random binary variable that decides which modality will contain the relevant information that determines the class. The input features of each modality can be generated by a random source, $\hat{y}_v$ and $\hat{y}_t$, or by an informed source, y_v and y_t. The generative model is specified as follows:

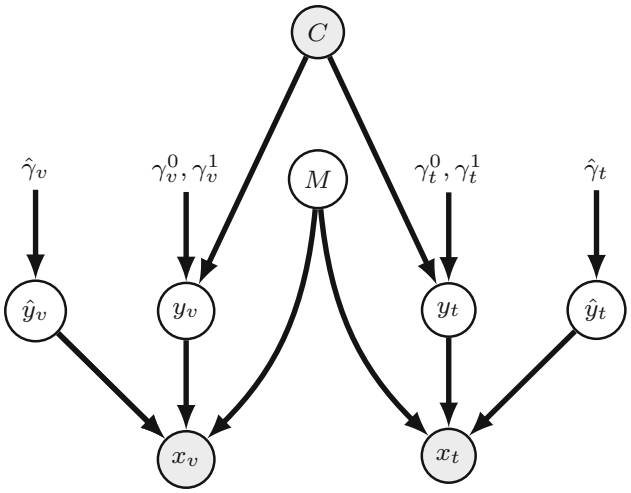


Fig. 3 Generative model for the synthetic task. Grayed nodes represent observed variables, the other nodes represent hidden variables. The goal is to estimate $P(C|x_v, y_v)$. M is a random binary variable that decides whether x_v or x_t will contain the relevant information that determines C. γ's are the parameters for noisy and relevant modalities

$$\begin{aligned}
C &\sim \text{Bernoulli}(p_C) \\
M &\sim \text{Bernoulli}(p_M) \\
y_v &\sim \mathcal{N}(\gamma_v^C) \\
\hat{y}_v &\sim \mathcal{N}(\hat{\gamma}_v) \\
x_v &= My_v + (1 - M)\hat{y}_v \\
y_t &\sim \mathcal{N}(\gamma_t^C) \\
\hat{y}_t &\sim \mathcal{N}(\hat{\gamma}_t) \\
x_t &= M\hat{y}_t + (1 - M)y_t
\end{aligned}$$

We trained a model with a single GMU and applied a sigmoid function over h, then the binary cross entropy (BCE) was used as loss function:

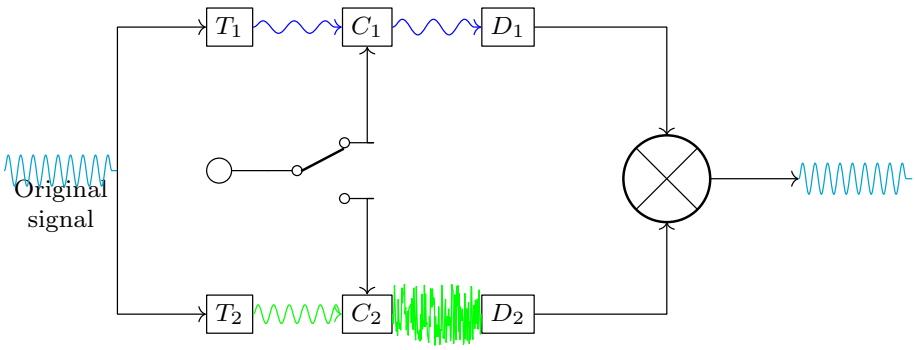


Fig. 2 Noisy channel model. There is an original source signal that is transformed by two independent components T_1 and T_2. The signals from T_1 and T_2 are transmitted by two channels, C_1 and C_2, respectively, that have two operation modes. In mode one, the channel transmits the original signal, in mode two, it transmits noise. A switch controls which channel will carry the signal. In one position, C_1 carries the signal and C_2 carries noise, in the other position, the situation is inverted. The switch may change its position at any time

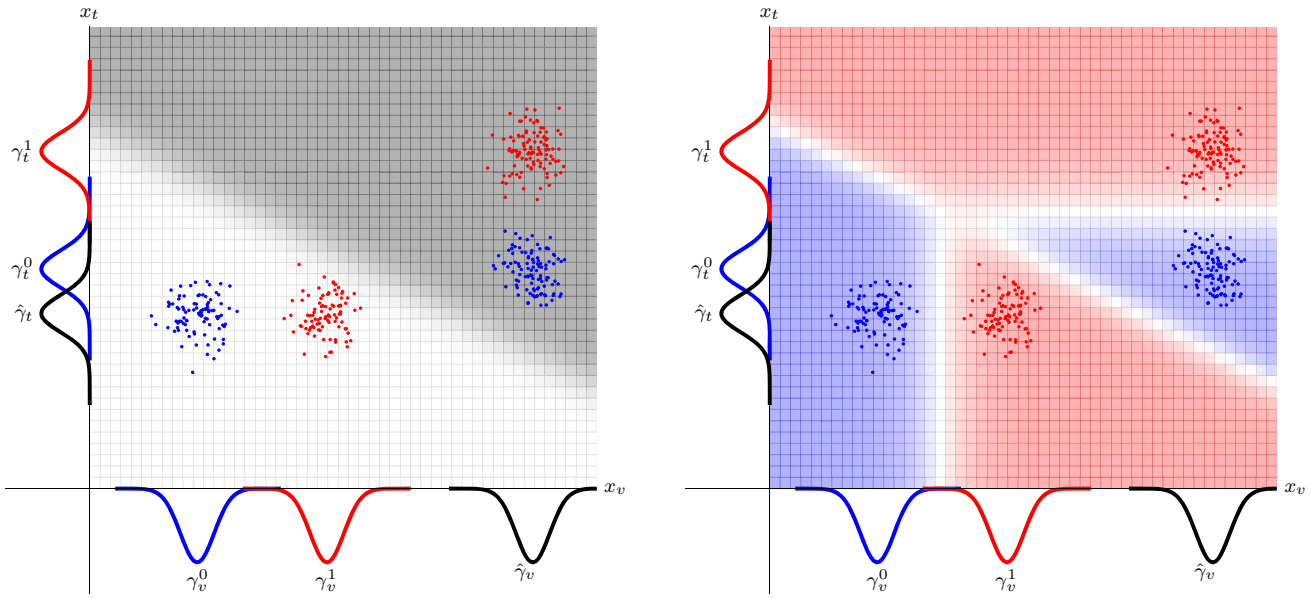


Fig. 4 Activations of z (left) and prediction (right) for a synthetic experiment with $x_v, x_t \in \mathbb{R}$ For a single bimodal GMU. Each axis represents a modality (for interpretation of the references to color in this figure legend, the reader is referred to the web version of this article)

$$\text{BCE}_{\text{synthetic}} = \sum_{i}^{n} c_i \log(\sigma(h_i)) + (1 - c_i) \log(1 - \sigma(h_i))$$

where σ is the sigmoid function, n is the number of samples and c_i the label for a given input pair $(x_v, x_t)_i$. Using the generative model, $n = 200$ samples per class were generated for each experiment. We ran 1000 synthetic experiments with different random seeds and the GMU outperformed a logistic regression classifier in 370 of them, while obtaining equal results in the remainder ones. Our goal in these simulations was to show that the model was able to learn a latent variable that determines which modality carries the useful information for the classification. An interesting result is that between M and the activations of the gate z there is a correlation of 1. This means the model was capable of learning such latent variable by only observing the x_v and x_t input features.

We also wanted to project back the z activations to the feature space in order to visualize regions depending on the modality. Figure 4 shows the activations in a synthetic experiment generated by the setup of Fig. 3 for $x_v, x_t \in \mathbb{R}$. Each axis represents a modality, red and blue dots are the samples generated for the two classes, and black Gaussian curves represent the $\hat{\gamma}_v$ and $\hat{\gamma}_t$ noises. The gray (white) regions of the left figure represent the activation of z. Notice that in the white region ($z = 1$), the model gives more importance to the x_v modality while in gray regions ($z = 0$) the x_t modality is more relevant; i.e., the z gate is isolating the noise. The contour of the right figure (blue–red) represents the model prediction. It is noteworthy that the boundary defined by the gates still holds when the model solves the task. This also encourages the inclusion of nonlinearities to the z gate so that it is able to discriminate more complex interactions between modalities.

4 Experiments

This section details the evaluation of GMU networks and other multimodal learning baseline methods for two supervised tasks: Genre classification (Sect. 4.1) using text and images, and image segmentation (Sect. 4.2) using RGB and depth images. Early fusion, late fusion and GMU were evaluated for both tasks.

4.1 GMU for genre classification

The Multimodal IMDb (**MM-IMDb**)[1] dataset [6] was built with the IMDb id's provided by the Movielens 20M dataset[2] that contains ratings of 25, 959 movies along with their plot, poster, genres and other 50 additional metadata fields such as year, language, writer, director, and aspect ratio. Notice that one movie may belong to more than one genre. Each plot contains on average 92.5 words, while the longest one contains 1431 words and the average of genres per movie is 2.48. In this work, we defined the task of movie genre prediction based on its plot and image poster. Nevertheless, the additional metadata information

[1] http://lisi1.unal.edu.co/mmimdb/.

[2] http://grouplens.org/datasets/movielens/.

encourages other interesting tasks such as rating prediction and content-based retrieval, among others.

The proposed model for genre classification is presented in Fig. 5. Both modalities are represented with pretrained models. Then the feature vectors are fused using the GMU. Finally a multilayer perceptron (MLP) with maxout units is stacked on top.

4.1.1 Data representation

Given that the nature of data for each modality is different, their statistical properties usually are not shared across modalities [59]. Thus, an evaluation of different strategies for representing visual and textual content are required. For text information, we evaluated word2vec models, *n*-grams models and RNN models. For processing visual data, we evaluated two different convolutional neural networks. The details of each representation are discussed below.

4.1.2 Text representation

Text representation is a critical step when classification tasks are addressed using machine learning methods. Traditional approaches are based on counting frequencies of *n*-gram occurrences such as words or sequences of characters (e.g., bag-of-words models). The main drawback of such approaches is the difficulty to model relationships between words and their context. An alternative approach was initially proposed by Bengio et al. [9], by building a neural network language model (NNLM). The NNLM was able to learn distributed representations of words that capture contextual information. Later, this model was simplified to deal with large corpora by removing hidden layers in the neural network architecture (word2vec) [51]. This is a fully unsupervised model that takes advantage of large sets of unlabeled documents. Herein, three text representations were evaluated:

N-grams Following the strategy proposed by Kanaris and Stamatatos [32], we used the character 3-gram strategy for representing text. Despite their simplicity, *n*-gram models have shown to be a competitive baseline.

Word2Vec Word2vec is an unsupervised learning algorithm that finds a vector representation for each word based on its context [51]. It has been shown that this model is able to find semantic and syntactic relationships using arithmetic operations between the vectors. Based on this property, we represent a movie as the average of the vectors of words in the plot outline. The main motivation to aggregate word2vec vectors is the property of additive compositionality that this representation has exposed over different sets of tasks such as word analogies. The usual way to aggregate the word vectors to represent a document is to perform arithmetic operations over the vectors. We take the average to avoid large input values to the neural network.

We used the pretrained Google Word2vec[3] embedding space composed by 300 dimensions. There were 41, 612 words from the MM-IMDb plots that are in the Google word2vec vocabulary. Other than lowercase, no text preprocessing nor stop word removal was applied. This textual representation obtained comparable state-of-the-art results [32] in two publicly available datasets: *7genre* dataset that comprises 1400 web pages with 7 disjoint genres and *ki-04* dataset that comprises 1239 samples classified under 8 genres. Notice that the state-of-the-art model [32] used character *n*-grams with structured information from the HTML tags to predict the genre of web pages, while ours only used the plain text.

Recurrent neural network This model takes as input a sequence of words to train a supervised recurrent neural network. Two variants were evaluated: (1) *RNN_w2v*, a transfer learning model that takes as input the word vectors of word2vec as representations; (2) *RNN_end2end*, which learns the word vectors from scratch.

4.1.3 Visual representation

In computer vision tasks, convolutional neural networks have become the de facto standard. It has been shown that CNN models trained with a huge amount of data are able to learn common features shared across different domains. This characteristic is usually exploited by transfer learning approaches. For visual representation, we explored 2 strategies: transfer learning and end-to-end training.

VGG Transfer VGG [56] is a neural network trained with the ImageNet dataset to classify natural images. We removed the classification layer from the VGG and propagated the images through it to get the last hidden layer activations as the visual representation.

End2End CNN Here, a CNN with 5 convolutional layers and an MLP (see Sect. 4.1.2) on top was trained from scratch. The first visual approach, *VGG_Transfer*, uses VGG network as feature extractor. The second approach takes as input the raw RGB images to a CNN. Since all the images do not have the same size, all images were scaled, and cropped when required, to 160×256 pixels keeping the aspect ratio. This CNN comprises 5 CNN layers of 5, 3, 3, 3, 3 squared filters and 2×2 pool sizes. Each convolutional layer has 16 hidden units. The convolutional layers are connected with the *MaxoutMLP* classifier on top.

[3] https://code.google.com/archive/p/word2vec/.

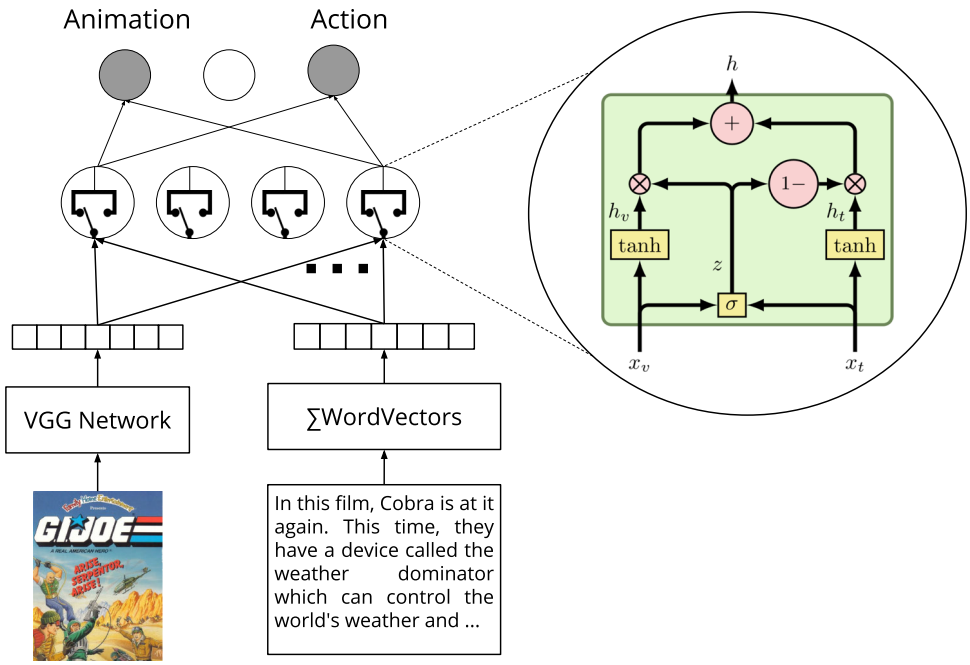


Fig. 5 Integration of the GMU in a multilayer perceptron for genre classification. Movie posters and movie plots were represented by the pretrained VGG network and the average of the Google word vectors, respectively. A layer with GMUs is used to combine both representations; finally, a multilabel classifier is stacked on top

4.1.4 Classification model

For classification stage, two methods to map from feature vectors to genre classification were explored: (1) Logistic regression and (2) a multilayer perceptron (MLP) with fully connected layers and maxout activation function. The maxout activation function $h_i : \mathbb{R}^n \rightarrow \mathbb{R}$ is a defined as:

$$h_i(\mathbf{s}) = \max_{j \in [1,k]} z_{i,j}$$

where $\mathbf{s} \in \mathbb{R}^n$ is the input vector, $z_{i,j} = \mathbf{s}^T \mathbf{W}_{\ldots ij} + \mathbf{b}_{ij}$ is the output of the jth linear transformation of the ith hidden unit, and $\mathbf{W} \in \mathbb{R}^{d \times m \times k}$ and $\mathbf{b} \in \mathbb{R}^{m \times k}$ are learned parameters. It has been shown that maxout models with just 2 hidden units behave as universal approximators, while are less prone to saturate units [22]. Since our intention is to measure how the network's depth affects the model performance, we evaluate the architecture with one and two fully connected layer.

4.1.5 Multimodal fusion baselines

We evaluate five different ways to combine both modalities as baselines.

Average probability This can be seen as a late-fusion strategy. The probabilities obtained by the best model of each modality are averaged and thresholded.

Concatenation Different works have found that a simple concatenation of representations of different modalities are good for combining the information [34, 55, 61]. Herein, we concatenated both representations to train the *MaxoutMLP* architecture.

Linear sum Following the way Vinyals et al. [67] combine text and images representation into a single space, this model adds a linear transformation for each modality so that both outputs have the same size to be summed up and then followed by the *MaxoutMLP* architecture.

MoE The mixture of experts (MoE) [29] model was adapted for multilabel classification. Two gating strategies were explored: *tied*, where a single gate multiplies all the logistics outputs, and *untied* where every logistic output has its own gate. Logistic regression and *MaxoutMLP* were evaluated as experts.

DeepCCA Deep canonical correlation analysis [4] is another way to perform information fusion. In this approach, the goal is to maximize the correlation between the modalities to later use the new correlated representation as input to the classifier.

4.1.6 Experimental setup

The MM-IMDb dataset has three subsets: Train, development and test subsets contain 15,552, 2608 and 7799, respectively. The sample was stratified so that training, dev and test sets comprise 60%, 10%, 30% samples of each genre, respectively.

In the multilabel classification, the performance evaluation can be more complex than traditional *multi-class* classification and the differences can be significant among several measures [46]. Herein, four averages of the *f*-score

(f_1) are reported: *samples* computes the *f*-score per sample and then averages the results, *micro* computes the *f*-score using all predictions at once, *macro* computes the *f*-score per genre and then averages the results. *weighted* is the same as *macro* with a weighted average based on the number of positive samples per genre. *F*-scores are calculated as follows [46]:

$$
\begin{aligned}
p^{\text{micro}} &= \frac{\sum_{j=1}^{Q} tp_j}{\sum_{j=1}^{Q} tp_j + \sum_{j=1}^{Q} fp_j} \, r^{\text{micro}} \\
&= \frac{\sum_{j=1}^{Q} tp_j}{\sum_{j=1}^{Q} tp_j + \sum_{j=1}^{Q} fn_j} \\
f_1^{\text{micro}} &= \frac{2 \times p^{\text{micro}} \times r^{\text{micro}}}{p^{\text{micro}} + r^{\text{micro}}} \\
f_1^{macro} &= \frac{1}{Q} \sum_{j=1}^{Q} \frac{2 \times p_j \times r_j}{p_j + r_j} \\
f_1^{\text{sample}} &= \frac{1}{N} \sum_{i=1}^{N} \frac{2 \times |\hat{y}_i \cap y_i|}{|\hat{y}_i| + |y_i|} f_1^{\text{weighted}} \\
&= \frac{1}{Q^2} \sum\nolimits_{j=1}^{Q} Q_j \frac{2 \times p_j \times r_j}{p_j + r_j}
\end{aligned}
$$

With N the number of examples; Q the number of labels; Q_j the number of true instances for the jth label; p the precision, r the recall; $\hat{y}_i, y_i \in (0,1)^Q$ the prediction and ground truth binary tuples, respectively; tp_j,fp_jandfn_j the number of true positives, false positives and false negatives for the jth label, respectively.

4.1.7 Neural network training

Neural network models were trained using batch normalization scheme [26]. This strategy applies a normalization step across samples that belong to the same batch, so that each hidden unit in the network receives a zero-mean and unit variance vector. Stochastic gradient descent with ADAM optimization [36] was used to learn the weights of the neural network. Dropout and max-norm regularization were used to control overfitting. Hidden size ($\{64, 128, 256, 512\}$), learning rate ($[10^{-3}, 10^{-1}]$), dropout ([0.3, 0.7]), max-norm ([5, 20]) and initialization ranges ($[10^{-3}, 10^{-1}]$) parameters were explored by training 25 models with random (uniform) hyperparameter initializations and the best was chosen according to validation performance. It has been reported that this strategy is preferable over grid search when training deep models [10]. All the implementation was carried on with the Blocks framework [66] and our source code is available.[4]

During the training process, we noticed that batch normalization considerably helped in terms of training time and convergence, resulting in less sensitivity to hyperparameters such as initialization ranges or learning rate. Also, dropout and max-norm regularization strategies helped to increase the performance at test time.

Overall, the neural network parameters were optimized using stochastic gradient descent. Notice that the learnable parameters in the GMU, $\{W_v, W_t, W_z\}$, are used in conjunction with differentiable operations. This allows to seamlessly integrate the GMU with automatic differentiation toolkits such as PyTorch or Tensorflow. In particular, the loss function for multilabel classification is defined as follows,

$$
\text{BCE}_{\text{mmimdb}} = \frac{1}{n} \sum_{i}^{n} \left|\left| c_i \odot \log\left(f_\Theta\left((x_v, x_t)_i\right)\right)\right|\right|_1
$$

where n is the number of movies, Θ is the set of learnable parameters, c_i and $\hat{y}_i = f_\Theta\left((x_v, x_t)_i\right)$ are vectors representing the ground truth and the f_Θ model predictions of the ith movie, respectively, $c_i, \hat{y}_i \in \mathbb{R}^k$, k is the number of genres, $\odot$ denotes the Hadamard product and $||\cdot||_1$ denotes the L_1-norm.

4.1.8 Results

The McNemar test is used to determine whether the differences between the GMU and the second best model have statistical evidence ($p < 0.01$). This test is preferable over other options because it presents low Type I error when the algorithms are computationally expensive and can be executed only once [12, 18].

Table 1 shows the results in the proposed dataset. For the textual modality, the best performance is obtained by the combination of the word2vec representation with an MLP classifier. The behavior of all representation methods is consistent across the performance measures. Learning from scratch the RNN model performed the worst. We hypothesize this has to do with the lack of data to learn meaningful relations among words. It has been shown that millions of words are required to train a model such as word2vec that is able to exploit common regularities between word co-occurrences.

For the visual modality, the usage of pretrained models works better than training the model from scratch. It seems it is still a small dataset to learn all the complexities of the posters. Now, comparing the performance independently per genre, as in Table 2, it is interesting to notice that in *Animation* the visual modality outperforms the textual one.

In the multimodal scenario, by adding the GMU as building block to learn the fusion we obtained the best performance, improving independent modalities in the

[4] https://github.com/johnarevalo/gmu-mmimdb.

Table 1 Summary of classification results on the MM-IMDb dataset

Modality	Representation	*F*-score			
		Weighted	Samples	Micro	Macro
Multimodal	GMU	**0.624**	**0.634**	**0.636**	**0.549**
	Linear_sum	0.606	0.617	0.617	0.520
	Concatenate	0.599	0.607	0.609	0.520
	AVG_probs	0.604	0.616	0.615	0.491
	MoE_MaxoutMLP	0.592	0.593	0.601	0.516
	MoE_MaxoutMLP (tied)	0.579	0.579	0.587	0.489
	MoE_Logistic	0.541	0.557	0.565	0.456
	MoE_Logistic (tied)	0.483	0.507	0.518	0.358
	DeepCCA	0.259	0.333	0.345	0.095
Text	MaxoutMLP_w2v	0.604	0.607	0.612	0.528
	RNN_transfer	0.570	0.580	0.580	0.480
	MaxoutMLP_w2v_1_hidden	0.540	0.540	0.550	0.440
	Logistic_w2v	0.530	0.540	0.550	0.420
	MaxoutMLP_3grams	0.510	0.510	0.520	0.420
	Logistic_3grams	0.510	0.520	0.530	0.400
	RNN_end2end	0.490	0.490	0.490	0.370
Visual	VGG_Transfer	0.416	0.436	0.449	0.284
	CNN_end2end	0.370	0.350	0.340	0.210

Text, visual and multimodal refers to models that used only the movie plot, only the movie poster and both modalities, respectively. Performance is reported using 4 types of *F*-score: micro-, macro-samples and weighted. Best results are shown in bold typeface

averaged measures and in 16 of out 23 genres and outperforming all other evaluated fusion strategies. The concatenation or the linear combination approaches were not enough to model the correlation between the modalities and MoE models did not perform better than simpler approaches. This is an expected behavior for MoE in a relatively small dataset because the data is fractionated over different experts, and thus it doesn't make an efficient use of the training samples. DeepCCA did not show promising results in this dataset. Since both modalities contain noisy samples, it seems DeepCCA forced to correlate information from one modality with the noise in the second one, downgrading the discriminative power, and thus the performance of single modality approaches.

GMU outperformed the best unimodal result. This result indicates that, unlike the other multimodal approaches, the GMU is not only robust to non-informative modalities in combination with a good one, but also manages to learn how to take advantage from the "noisy" modality to improve classification in some cases. The remainder models were not able to take advantage of the non-informative modality.

Results for the McNemar test are presented in Table 2. Statistical evidence showed that there is a significant difference between the GMU and the second best model for *Drama*, *Crime*, *Action*, *Horror*, *Family*, and *Short* genres. In the remainder genres the GMU shares the first place with another method.

When designing the network architecture, we explored the number of required layers in the MLP classifier, finding that 2 fully connected layers were sufficient to achieve the best performance. On the GMU side, since the unit does not contain multiple components, the only operation that could be removed is the gating mechanism. By doing so the fusion would be equivalent to the *LinearSum* baseline. We also explored ReLU ($f(s) = \max(0, s)$) as nonlinearity for the h_v and h_t activations, finding negligible variations in the final performance.

In order to evaluate which modality influences the model more when assigning a particular label, we averaged the activations of a subset of z gates of the test samples to which the model assigned them such label. We counted the number of samples that pays more attention to the textual modality ($z \leq 0.5$) or to the visual modality ($z > 0.5$). The units were chosen taking into account the mutual information between the predictions and the z activations. The result of this analysis is depicted in Fig. 6. As expected, the model is generally more influenced by the textual modality. But, in some specific genres such as *Animation* or *Family*, the visual modality affects more the model. This is also consistent with results of Table 2 that reports better performances for visual modality.

Table 2 Macro *F*-score reported per genre for single and multimodal approaches

Genre	Textual	Visual	GMU
Drama	0.75	0.68	**0.77***
Comedy	0.63	0.59	**0.67**
Romance	0.52	0.32	**0.52**
Thriller	0.58	0.41	**0.61**
Crime	0.63	0.27	**0.65***
Action	0.58	0.38	**0.62***
Adventure	**0.53**	0.32	0.51
Horror	0.65	0.43	**0.70***
Documentary	0.75	0.18	**0.76**
Mystery	0.39	0.12	**0.39**
Sci-Fi	0.66	0.31	**0.67**
Fantasy	**0.45**	0.22	0.44
Family	0.51	0.47	**0.58***
Biography	**0.40**	0.01	0.25
War	0.65	0.16	**0.66**
History	**0.41**	0.06	0.37
Music	0.57	0.04	**0.57**
Animation	0.43	0.61	**0.65**
Musical	0.22	0.19	**0.27**
Western	0.64	0.33	**0.68**
Sport	**0.69**	0.14	0.68
Short	0.29	0.20	**0.30***
Film-Noir	0.20	0.09	**0.30**

Best results are shown in bold typeface

*Marks scores with evidence of difference between the GMU and the second best model ($p < 0.01$)

Notice that the proposed model outperformed 13 out of 18 genres and obtained competitive results in *Adventure*, *Fantasy History* and *Sport* genres. *Biography* genre performed the worst in both GMU and visual-only models. We hypothesize this behavior is expected due to the noisy input in the images. Figure 7 illustrates posters in which could be challeging, even for a human, to associate it with the *Biography* genre. In these scenarios the GMU may require more data to better learn to filter such noisy modalities. Overall, in terms of macro-performance measures the proposed model outperformed single and multimodal baselines.

We did a qualitative analysis of instances where performance improved by a relatively large margin. Table 3 illustrates cases where the model takes advantage of the most accurate modality, and in some cases removes false positives. It is noteworthy that some of these examples can be confusing for a human if one modality is missing, or additional information is not given.

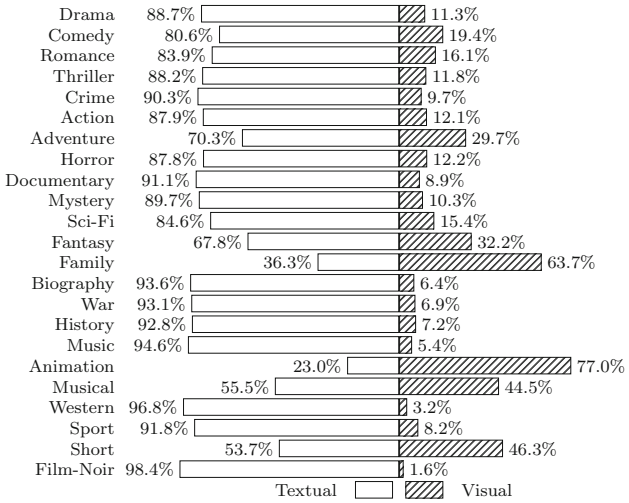


Fig. 6 Percentage of gates activations ($z > 0.5$: visual; $z \leq 0.5$: textual) per modality for a subset of the GMUs. The units were chosen using the mutual information between the predictions and the z activations. All predictions for each genre in the test set were used in this analysis (i.e., correct and incorrect predictions)

4.2 GMU for image segmentation

The proposed unit is easily adaptable to other architectures different from the traditional "Fully connected". Since the GMU is a differentiable operator, it can be applied to part of the input and still be optimized with gradient-based methods. This is the basic idea of convolutional architectures. This section adapts the GMU to convolutional neural networks for addressing image segmentation. The model learns to fuse RGB and depth information to outperform standard early- and late-fusion strategies. An analysis to the learned model highlights correlations between modalities and semantic concepts.

Convolutional architectures are widely used in image processing scenarios. Shortly after the Imagenet success [39], CNN became the de facto standard architecture when using neural networks for image representation. In CNN, there are convolution and pooling transformations to the input image. Consider the input image $M \in \mathbb{R}^{p \times p}$, the first transformation applies a convolution with a filter $K \in \mathbb{R}^{k \times k}$ to obtain a feature map $S \in \mathbb{R}^{(p-k+1)\times(p-k+1)}$, followed by a nonlinearity activation function $a : \mathbb{R} \Rightarrow \mathbb{R}$ applied in an element-wise fashion. The second transformation reduces the dimension of the feature map by applying a local subsampling function over the output feature map $a(S)$.

4.2.1 Deep scene dataset

The convolutional architecture is evaluated in the DeepScene dataset[5] [65]. The dataset was collected using an

[5] http://deepscene.cs.uni-freiburg.de/.

Fig. 7 Posters picked from the test set with *Biography* genre. For most of these movies, their visual appearance is not sufficient to determine the topic

autonomous mobile robot platform equipped with a stereo vision camera and a modified dashcam for acquiring RGB and near-Infrared (NIR) data, respectively. Both cameras were time synchronized and frames were captured at 20Hz. Additional image post-processing was applied to match both images. Figure 8 shows the autonomous robot platform and one example with the available modalities.

The data was collected on three different days to have variability in lighting conditions as shadows and sun angles play a crucial role in the quality of acquired images. The DeepScene dataset comprises 366 images with pixel-level ground-truth annotations which were manually annotated with 1 out of the 6 concepts: {*grass*, *obstacle*, *tree*, *vegetation*, *road* and *sky*}. It also provides train and test sets with 230 and 135 scenes, respectively.[6]

DeepScene dataset authors also computed Global-based vegetation indices such as Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) to extract consistent spatial and global information [25]. Depth images were obtained using the approach from Liu et al. [42] that employs a deep convolutional neural field model for depth estimation by constructing unary and pairwise potentials of conditional random fields. The Multispectrum channel fusion NRG (near-Infrared, red, green) image was also computed and included as another modality. We choose RGB and Depth images as input to the proposed multimodal approach because these are the most common and general modalities. The remainder ones are specific for environments with abundant presence of vegetation.

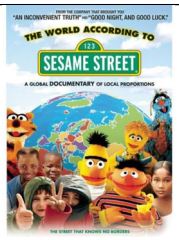
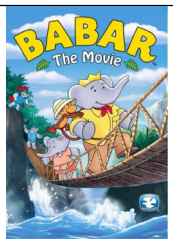
4.2.2 Convolutional GMU for segmentation

Some tasks involve multimodal sources that are suitable to be represented by a convolutional architecture. This is the case of image segmentation using RGB and depth images. Both of them represent the same scene, but using different information. Also, both of them can be naturally represented by a CNN. This is a convenient scenario to apply the GMU and let the model learn which parts of the image are more relevant to the classification.

We integrated the GMU in a convolutional architecture as depicted in Fig. 9, where the GMU layer takes S_{rgb} and S_{depth} feature maps as inputs and outputs a combined feature map. A convolutional neural network is used to learn directly from the pixels the representation for each modality. Convolutional filters are 3×3 with padding of 1, except the last convolutional layer which has a kernel size 4×4 with zero-padding. Then a set of deconvolutional layers are stacked to reconstruct the original resolution of

[6] We discarded the image with ID *b275-311* from test set because it is incorrectly annotated.

Table 3 Qualitative evaluation of predictions in test set

The World According to Sesame Street	
	a documentary which examines the creation and co - production of the popular children's television program in three developing countries: bangladesh,kosovo and south africa
Ground Truth	Documentary
Textual	**Documentary**, *History*
Visual	**Comedy**, *Adventure*, *Family*, *Animation*
GMU	**Documentary**
Letters from Iwo Jima	
	the island of iwo jima stands between the american military force and the home islands of japan. (...) when the american invasion begins,both kuribayashi and saigo find strength, honor,courage,and horrors beyond imagination
Ground Truth	Drama, War, History
Textual	**Drama**, *Action*, **War**, **History**
Visual	*Thriller*, *Action*, *Adventure*, *Sci-Fi*
GMU	**Drama**, **War**, **History**
Babar: the movie	
	in his spectacular film debut,young babar,king of the elephants,must save his homeland from certain destruction by rataxes and his band of invading rhinos
Ground Truth	Adventure, Fantasy, Family, Animation, Music
Textual	**Adventure**, *Documentary*, *War*, *Music*
Visual	*Comedy*, **Adventure**, **Family**, **Animation**
GMU	**Adventure**, **Family**, **Animation**

Bold and italic genres are true positives and false positives respectively. In these examples, the GMU model was able to take advantage of both modalities to remove false positives and to include non-predicted genres by the single-modality approaches

the image. A convolutional GMU layer is applied to fuse both RGB and depth feature maps. Finally, a softmax layer is stacked on top. Notice that parameters of both convolutional networks are shared.

Let $g : \mathbb{R}^{300\times300\times3} \rightarrow \mathbb{R}^{300\times300\times32}$ be the function representing the convnet depicted in Fig. 9. This convnet is applied to each input modality to get $S_{\text{rgb}}, S_{\text{depth}}$ features maps. Then, the convolutional GMU is applied as follows:

$$
\begin{aligned}
h_{\text{rgb}} &= \tanh\left(W_{\text{rgb}} * S_{\text{rgb}}\right) \\
h_{\text{depth}} &= \tanh\left(W_{\text{depth}} * S_{\text{depth}}\right) \\
z &= \sigma\left(W_z * \left[S_{\text{rgb}}, S_{\text{depth}}\right]\right) \\
h &= z \odot h_v + (1 - z) \odot h_t
\end{aligned}
$$

where $*$ denotes the convolutional operator, $\odot$ denotes the Hadamard product, $[\cdot,\cdot]$ the concatenation operator, $h \in \mathbb{R}^{300\times300\times q}$ is the fused representation and q the number of hidden units in the GMU layer, which is later used as input to the softmax classifier.

4.2.3 Experimental setup

We took 46 scenes from train as our validation set to tune hyperparameters of the model. Hyperparameters were explored by training 25 models with random (uniform) hyperparameter initializations and the best was chosen according to validation performance.

Following the dataset authors' approach [65], images were preprocessed by resizing the original image to 300×300 pixels keeping the aspect ratio and cropping them when necessary. During training, images were oversampled by applying random rotations between $[-30, 30]$ degrees, random flipping and random cropping the images. Previous works [39] have reported this as a convenient way to artificially increase the number of training samples, which in turn helps to better generalization during the model training.

The convolutional architecture used in these experiments is detailed in Fig. 9. The pixel-based classification

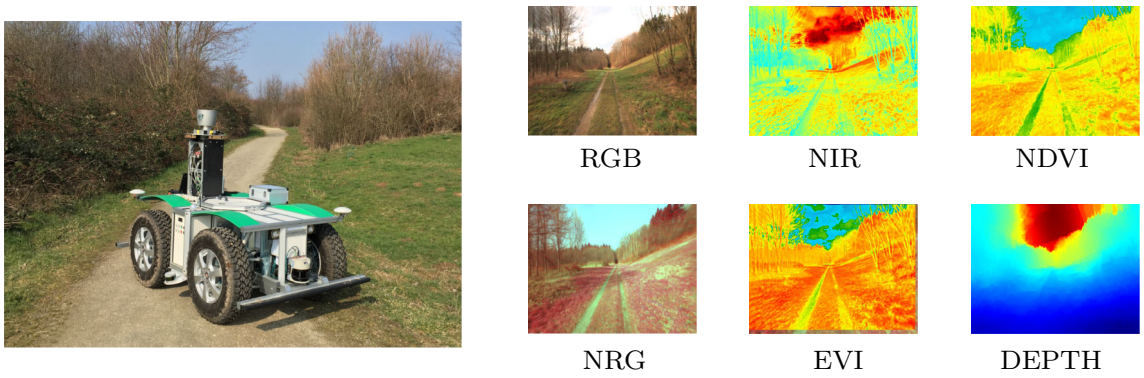


Fig. 8 Left: Robot used to capture the images. Right: sample from the Deep scene dataset with the available modalities. Acquisition involved two cameras for RGB and near-Infrared (NIR) images. Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), Multispectrum channel fusion NRG (near-Infrared, red, green) and depth images were computed based on the RGB and the NIR images. (Image taken from Valada et al. [65])

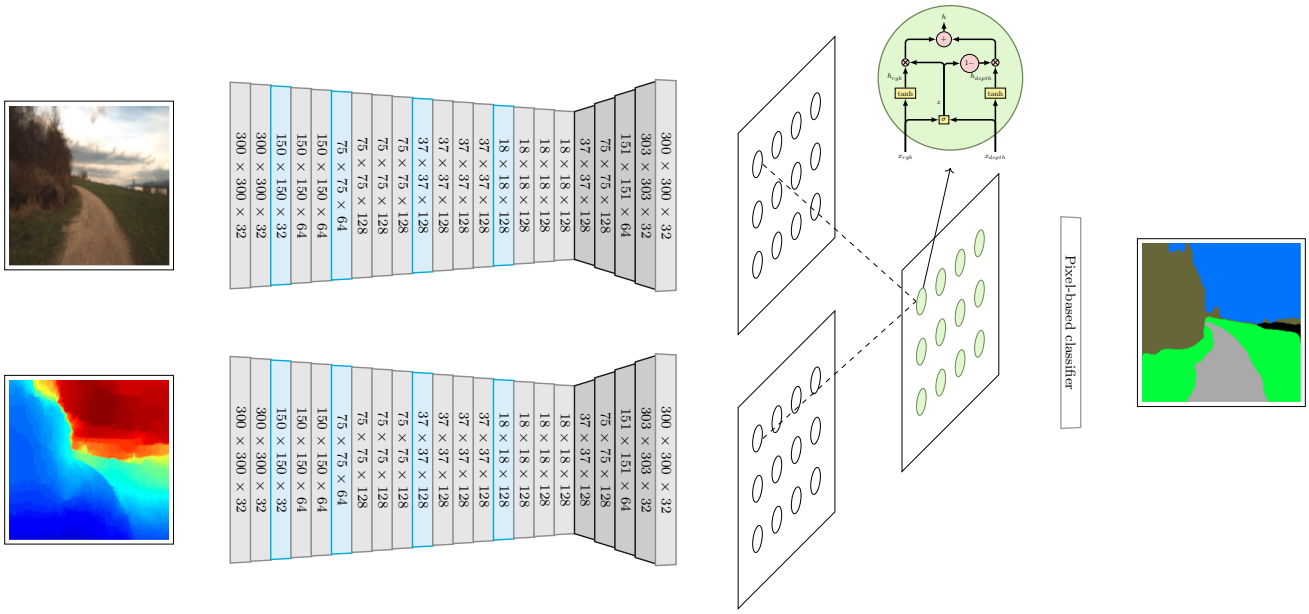


Fig. 9 Integration of the proposed GMU in convolutional architectures. Light gray, dark gray and cyan represent convolutional, deconvolutional and pooling layers, respectively. Output dimensions are denoted inside each layer. Convolutional filters are 3×3 with padding of 1, except the last convolutional layer which has a kernel size 4×4 with zero-padding. Parameters of both convolutional networks are shared (colour figure online)

layer after this deep model varies depending on the model used. For single-modality approaches, the last layer is a convolution with 6 filters of 3×3 with padding of 1 to keep the 300×300 size followed by a Softmax activation function. For the multimodal approach there is an additional layer with 32 filters for each modality, then the ConvGMU layer that merges those 32 pairs of feature maps, followed by a Softmax activation layer. Experiments are supported by the McNemar statistical test to determine whether the differences have statistical evidence ($p < 0.01$).

4.2.4 Baseline

A natural way to perform the fusion in a convolutional architecture is to stack the depth image as a fourth channel in the RGB image. We named it as the *concatenation* approach. Similarly to the IMDb setup, we also included the *LinearSum* and the *AvgProb* as early and late fusions, respectively. In the *LinearSum* approach we applied an element-wise summation over the last feature map of each modality, the resultant map is the input for the Softmax classifier. In the *AvgProb* approach, the probability maps are averaged in a pixel-wise manner.

4.2.5 Results

Firstly, it is noteworthy that some inconsistencies in the ground truth were highlighted when visualizing the predictions. Figure 10 depicts that *obstacle* and *tree* concepts are correctly annotated in training set, but are wrongly annotated in the test set. Due to such inconsistencies, in this experimentation those two concepts were discarded when methods were compared.

Following the original paper, Intersection over union (IoU), accuracy (ACC), false positive rate (FPR) and false negative rate (FNR) are used as performance measures. Table 4 summarizes the results for unimodal and multimodal approaches. Results showed RGB outperformed the depth modality for all classes. Also, the behavior of other multimodal approaches is consistent with the results for the MM-IMDb dataset. Here, again the GMU approach outperformed both unimodal and multimodal methods. We applied the McNemar statistical test for paired data in a pixel-wise manner. The statistical evidence showed that the differences between GMU and the remainder models are significant ($p < 0.01$) for all the classes.

As noted in Table 5, IoU of *road* and *sky* concepts increased the most with the convolutional GMU model. This is consistent with the nature of the data, since closest and farthest concepts are closely related with the kind of information that depth images provide.

Likewise in the MM-IMDb task, an analysis of z activations with respect to the predictions is reported in Fig. 11. For *road*, *grass* and *vegetation* the RGB modality is more dominant. In contrast, for *tree*, *sky* and *obstacle* the depth modality gives more information for the classifier. We believe this is consistent with the nature of the data, since concepts such as *sky* and *obstacle* would be easier to detect when additional information like distance to camera is provided.

4.3 Discussion

Our results show that the GMU is a feasible multimodal fusion strategy to boost the performance in different neural network architectures. This improvement has been consistently supported by the scientific community: Yan and Zhao [74] and Yao et al. [73] integrated the GMU in a recurrent architecture to generate short texts for conversational systems. Kiela et al. [35] proposed a simplification of the GMU by tying the weights of the gate. Fernando et al. [20] integrated the GMU in a memory network architecture for pedestrian trajectory prediction. In [44], the GMU was used as baseline to introduce a new task called multimodal attribute extraction and Ye et al. [75] used the GMU as fusion module to grade glioma in magnetic resonance images.

In the previously mentioned scenarios the application of the GMU assumes the multimodal input is fully paired, i.e., each sample is a tuple of multiple representations of the same object. Notice this pairing is not always available and thus the immediate application of the GMU is not possible. There exists weakly-paired approaches such as [43, 49] which deal with input modalities as group of multiple objects.

In contrast, the GMU is intended to deal with noisy inputs as motivated in the synthetic scenario. Such noisy scenario is also present in the movies experiment where either the image or the text could be non-informative to describe the genre of the movie. Qualitative examples in Table 3 depict how this phenomenon affects the predictions of single-modality models and the benefit of using the GMU. Notice that in such cases, the modalities are not important alike to detect the movie genre, so for some instances one modality is more informative than another.

For all the problems explored in this work, we used the same strategy for initialization: random sampling from independent uniform distributions in the $[-\delta, \delta]$ range, being δ a hyperparameter explored in the range of $[10^{-3}, 10^{-1}]$. Similarly, the learning rate was explored in the $[10^{-5}, 10^{-2}]$ range, with Adam [36] as the default optimizer. Similarly to the baseline methods, experiments showed that 400 is a sufficient number of epochs for convergence of GMU networks. As consequence, it is reasonable to conclude that, regarding optimization and hyperparameter exploration, the GMUs showed the common behavior of other gradient-based models during the training process [8].

The experimental evaluation for all architectures herein implemented did not have troubles achieving perfect accuracy in the training set while obtaining a close-to-zero value in the training loss function. This supports the findings of Choromanska et al. [15] where it is stated that despite recovering the global minimum is harder in neural networks with large sizes, in practice is irrelevant as global minimum often leads to overfitting. Thus, as in others deep learning models, the challenge is not the convergence but to achieve a low generalization error. In this matter we used two regularization strategies: dropout and max-norm regularization. We did not find any particular consideration on the application of such regularizers when they were integrated with GMU networks.

The batch size mainly affects the convergence time, but not the performance. For vectorized implementations (e.g., linear algebra libraries or GPU) The bigger the batch size the shorter the training time. In particular, we set the batch size so that maximizes the usage of the GPU memory. That is, for small input size like word2vec or VGG vector representations we used batch sizes of 128 samples, for larger

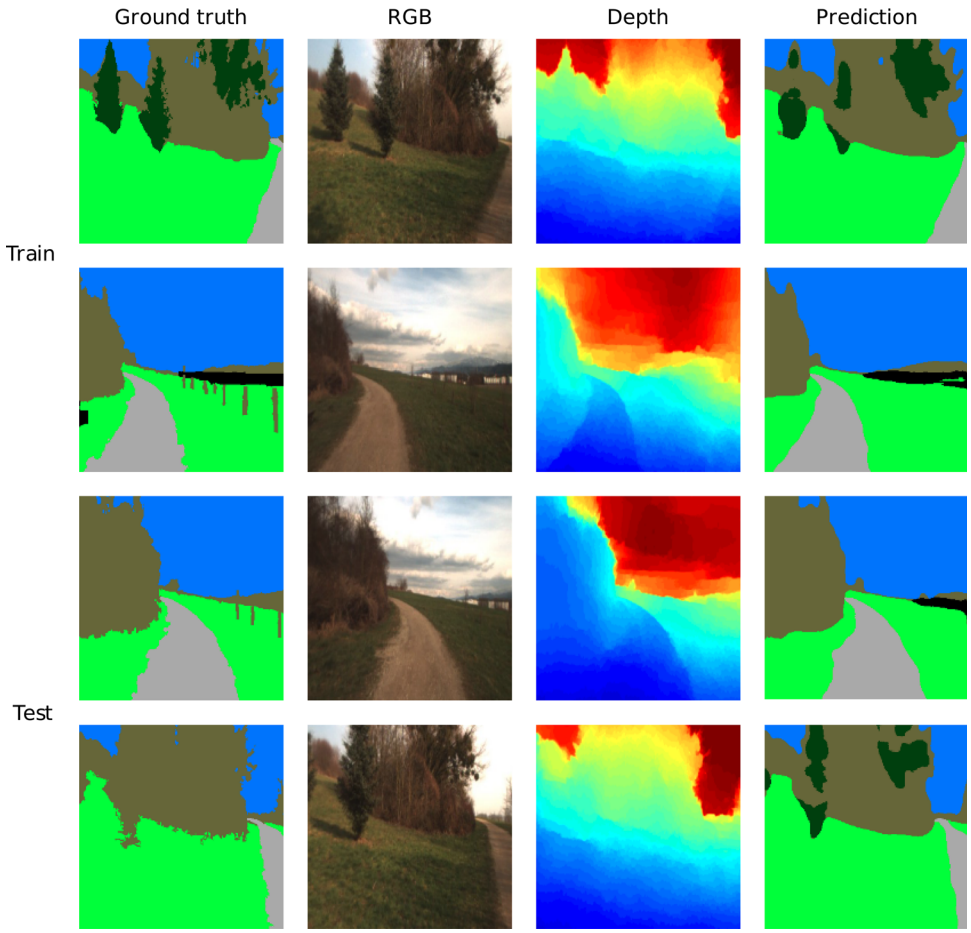


Fig. 10 Segmentation results for the convolutional network with GMU. Concepts for the first (ground truth) and fourth (prediction) columns are colored as follows: sky: blue, grass: light green, vegetation: olive, road: light gray, obstacle: black, tree: dark green. Note that obstacle and tree concepts are correctly annotated in the training set, but at test set are absent. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article) (colour figure online)

Table 4 Summary of image segmentation results using single (RGB and Depth) and multimodal (AvgProb, Concatenate, LinearSum and ConvGMU) approaches in the test set

Method	IoU	ACC	FPR	FNR
RGB	0.840	0.964	0.029	0.083
Depth	0.630	0.914	0.064	0.239
AvgProb	0.818	0.964	0.028	0.097
Concatenate	0.851	0.969	0.025	0.084
LinearSum	0.855	0.970	0.025	0.082
ConvGMU	**0.861**	**0.971**	**0.022**	**0.077**

Intersection over union (IoU), accuracy (ACC), false positive rate (FPR) and false negative rate (FNR) are reported. Best results are shown in bold typeface

Table 5 Intersection over union per class for unimodal (RGB and depth) and multimodal (ConvGMU) approaches

Method	Road	Grass	Vegetation	Sky
RGB	0.784	0.822	0.891	0.863
Depth	0.392	0.574	0.774	0.780
ConvGMU	**0.828**	**0.842**	**0.893**	**0.880**

Best results are shown in bold typeface

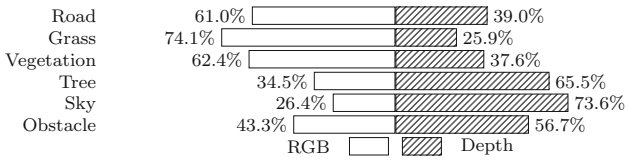


Fig. 11 Percentage of gates activations in the image segmentation task ($z > 0.5$: RGB; $z \leq 0.5$: depth) per modality for a subset of the GMUs. The units were chosen using the mutual information between the predictions and the z activations. All predictions for each concept in the test set were used in this analysis (i.e., correct and incorrect predictions)

input sizes like two images in the DeepScene dataset, the GPU was able to host up to 32 samples in each batch.

The GMU is closely related to other architectures that use gates to control the information flow such as LSTMs, Mixture of experts and attention mechanisms. GMU is related to the mixture-of-experts (MoE) model in the usage of multiplicative gates to control the information flow.

However the MoE includes gates after the inference step of multiple predictors. This differs from GMU in two aspects: (1) the MoE explicitly requires supervision to be used, whilst GMU can be used also in unsupervised approaches, and (2) In the MoE the data is ultimately fractionated over different experts, and thus it doesn't make an efficient use of the training samples in a relatively small dataset. This was evidenced in the results where MoE models did not perform better than simpler unimodal approaches, while GMU took advantage of both modalities to boost the performance.

The GMU design is influenced by the way that gated recurrent networks, such as LSTMs and GRUs, work. Gated recurrent networks process input sequences by controlling the information flow using multiplicative operations, which is the same approach followed by the GMU. The difference lies in the assumption of temporal dependencies in the input, GMU is not intended to deal with temporal or sequential data. Instead, it combines multiple inputs in a single operation. It is noteworthy, that both models can be integrated. Indeed, Yao et al. [74] used our GMU to fuse the hidden states of two recurrent networks.

Highway networks [60] are also gated-based models that control the information flow using sigmoid activation functions. The difference with respect to the GMU is twofold: (1) the GMU expects multiple inputs, corresponding to multiple modalities, to be merged while highway networks expect a single input, calculate the hidden representation of the merged inputs and then merge it with the original input, and (2) the gate function only attends the original input, while the GMU is able to attend all of the multimodal inputs.

Attention-based mechanisms have been proposed in other multimodal problems such as visual question answering (VQA). In VQA, the system receives an image and a text-free question statement related to the input image. The goal is to generate a text-free answer for the question. Wu et al. [71] built a comprehensive review on VQA research classifying the recent approaches in four categories: compositional models, knowledge-based enhanced, joint embedding and attention mechanisms. Compositional models aim to divide the problem in multiple tasks such as to question parsing, image representation, and answer generation. Some models includes an end-to-end approach, while others use isolated tools/methods. Knowledge-based enhanced approaches uses external databases to add priors that ease question parsing or answer generation steps. Joint embedding and attention models are mostly based in end-to-end approaches. Joint embedding learns a common feature spaces for representing both modalities. Attention mechanisms enhance the joint embeddings by focusing in specific parts of the image and the question. The GMU differs from these approaches in that it is agnostic to the task, i.e., it can be used in any neural network that involves a fusion step. GMU also is not assuming a common space across modalities, in contrast it maps their inputs to a new feature space built by a weighted combination of multimodal features.

Chen et al. [13] proposed an attention mechanism for image segmentation. They fused the output probability map of multiple classifiers, each one feed with a different scaled version of the input image. In order to train the model with gradient-based optimizers, its architecture requires differentiable bilinear interpolations to have the same output size for all the scales. Notice that this is equivalent to a mixture of experts model across multi-scale classifiers. The parallel between GMU and MoE models was previously discussed in the GMU definition in Sect. 3 and the performance analysis in Sect. 4.

5 Conclusions

This work presented a strategy to learn fusion transformations from multimodal sources. Similarly to the way recurrent models control the information flow, the proposed model is based on multiplicative gates. The Gated Multimodal Unit (GMU) receives two or more input sources and learns to determine how much each input modality affects the unit activation. This contrasts the traditional fusion methods that adjust weights for each modality and are fixed for all instances, while the GMU weights are determined by the input. In synthetic experiments the GMU was able to learn hidden latent variables, and in two real scenarios it outperformed the single-modality, early- and late-fusion approaches. A key property of the GMU is that, being a differentiable operation, it was easily coupled in different neural network architectures and trained with standard gradient-based optimization algorithms. The model was integrated with convolutional and fully connected networks for two different supervised tasks. In the movie genre classification task, the gated multimodal network involved a fully connected architecture taking as input the plot of the movie and the image poster to annotate (multilabel) 23 genres. Experimental evaluation showed the model learned to weight the modalities based on the input features, and outperformed early- and late-fusion approaches by 3% in terms of F-score. In the image segmentation task, the gated multimodal network involved an end-to-end convolutional architecture taking as input the RGB and depth images and output the segmented image with 6 semantic concepts. Likewise, the model outperformed other single and multimodal approaches measuring the Intersection-over-union score. The activations of the GMU layer were mapped to

the output concepts finding correlations between input modalities and output concepts, e.g., depth information was more correlated with "sky" and "tree" while RGB is more correlated with "grass" and "vegetation". It should be noted that even though the model is capable of combining information, the content representation is critical to correctly take advantage of the different modalities.

Interesting directions for future work include more complex transformations in the gate. As shown in the synthetic experiment, the proposed gate learned a linearity function to determine which modality has more information. This can be extended so that the gate applies nonlinear transformations to increase its flexibility. Another open challenge in multimodal representation learning is to deal with missing modalities. One alternative is to include a new learnable initial state for each modality. At test time, such state would be used as default value for absent modalities. The inclusion of other generative models such as adversarial networks are also interesting paths to deal with missing data.

Acknowledgements Arevalo thanks Colciencias for its support through a doctoral Grant in call 617/2013. This research was partially funded by CONACYT Project FC-2016/2410.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Akata Z, Lee H, Schiele B (2014) Zero-shot learning with structured embeddings. CoRR abs/1409.8. arxiv:1409.8403
2. Alvear-Sandoval RF, Figueiras-Vidal AR (2018) On building ensembles of stacked denoising auto-encoding classifiers and their further improvement. Inf Fusion 39:41–52
3. Anand D (2014) Evaluating folksonomy information sources for genre prediction. In: Advance computing conference (IACC), 2014 IEEE international, pp 887–892. https://doi.org/10.1109/IAdCC.2014.6779440
4. Andrew G, Arora R, Bilmes JA, Livescu K (2013) Deep canonical correlation analysis. In: ICML (3), pp 1247–1255
5. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Lawrence Zitnick C, Parikh D (2015) Vqa: visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp 2425–2433
6. Arevalo J, Solorio T, Montes-y Gómez M, González FA (2017) Gated multimodal units for information fusion. In: 5th international conference on learning representations 2017 workshop
7. Atrey PK, Hossain MA, El Saddik A, Kankanhalli MS (2010) Multimodal fusion for multimedia analysis: a survey. Multimed Syst 16(6):345–379. https://doi.org/10.1007/s00530-010-0182-0
8. Bengio Y (2012) Practical recommendations for gradient-based training of deep architectures. In: Montavon G, Orr GB, Müller KR (eds) Neural networks: tricks of the trade. Springer, Berlin, pp 437–478
9. Bengio Y, Ducharme R, Vincent P, Janvin C (2003) A neural probabilistic language model. J Mach Learn Res 3:1137–1155
10. Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. J Mach Learn Res 13(Feb):281–305
11. Bhatt C, Kankanhalli M (2011) Multimedia data mining: state of the art and challenges. Multimed Tools Appl 51(1):35–76. https://doi.org/10.1007/s11042-010-0645-5
12. Bouckaert RR, Frank E (2004) Evaluating the replicability of significance tests for comparing learning algorithms. In: Pacific-Asia conference on knowledge discovery and data mining. Springer, pp 3–12
13. Chen LC, Yang Y, Wang J, Xu W, Yuille AL (2016) Attention to scale: scale-aware semantic image segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3640–3649
14. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:14061078
15. Choromanska A, Henaff M, Mathieu M, Arous GB, LeCun Y (2015) The loss surfaces of multilayer networks. J Mach Learn Res 38:192–204
16. Coates A, Ng AY (2011) The importance of encoding versus training with sparse coding and vector quantization. In: Proceedings of the 28th international conference on machine learning (ICML-11), pp 921–928
17. Deng L (2014) A tutorial survey of architectures, algorithms, and applications for deep learning. APSIPA Trans Signal Inf Process. https://doi.org/10.1017/atsip.2013.9
18. Dietterich TG (1998) Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 10(7):1895–1923
19. Feng F, Li R, Wang X (2013) Constructing hierarchical image-tags bimodal representations for word tags alternative choice. arXiv preprint arXiv:13071275
20. Fernando T, Denman S, Sridharan S, Fookes C (2018) Pedestrian trajectory prediction with structured memory hierarchies. arXiv preprint arXiv:180708381
21. Frome A, Corrado GS, Shlens J, Bengio S, Dean J, Ranzato MA, Mikolov T (2013) DeViSE: a deep visual-semantic embedding model. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds) Advances in neural information processing systems, vol 26. Curran Associates Inc., Hook, pp 2121–2129
22. Goodfellow I, Warde-farley D, Mirza M, Courville A, Bengio Y (2013) Maxout networks. In: Dasgupta S, Mcallester D (eds) Proceedings of the 30th international conference on machine learning (ICML-13), JMLR workshop and conference proceedings, vol 28, pp 1319–1327
23. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
24. Huang EH, Socher R, Manning CD, Ng A (2012) Improving word representations via global context and multiple word prototypes. In: Proceedings of the 50th annual meeting of the association for computational linguistics: long papers, vol 1. Association for Computational Linguistics, pp 873–882
25. Huete A, Justice C, Van Leeuwen W (1999) Modis vegetation index (mod13). Algorithm Theor basis Doc 3:213
26. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of The 32nd international conference on machine learning, pp 448–456
27. Ivasic-Kos M, Pobar M, Mikec L (2014) Movie posters classification into genres based on low-level features. In: 2014 37th international convention on information and communication technology, electronics and microelectronics (MIPRO), vol i.

IEEE, pp 1198–1203. https://doi.org/10.1109/MIPRO.2014.6859750
28. Ivasic-Kos M, Pobar M, Ipsic I (2015) Automatic movie posters classification into genres. In: Bogdanova MA, Gjorgjevikj D (eds) ICT Innovations 2014: world of data. Springer International Publishing, Cham, pp 319–328. https://doi.org/10.1007/978-3-319-09879-1_32
29. Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE (1991) Adaptive mixtures of local experts. Neural Comput 3(1):79–87
30. Janowczyk A, Madabhushi A (2016) Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. J Pathol Inform 7:1–29
31. Johnson J, Karpathy A, Fei-Fei L (2015) Densecap: fully convolutional localization networks for dense captioning. arXiv preprint arXiv:151107571
32. Kanaris I, Stamatatos E (2009) Learning to recognize webpage genres. Inf Process Manag 45(5):499–512. https://doi.org/10.1016/j.ipm.2009.05.003
33. Kang Y, Kim S, Choi S (2012) Deep learning to hash with multiple representations. In: 2012 IEEE 12th international conference on data mining. IEEE, pp 930–935
34. Kiela D, Bottou L (2014) Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP-14), pp 36–45
35. Kiela D, Grave E, Joulin A, Mikolov T (2018) Efficient large-scale multi-modal classification. arXiv preprint arXiv:180202892
36. Kingma D, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:14126980
37. Kiros R, Salakhutdinov R, Zemel RS (2014a) Multimodal neural language models. ICML 14:595–603
38. Kiros R, Salakhutdinov R, Zemel RS (2014b) Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:14112539
39. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in neural information processing systems 25. Curran Associates Inc, New york, pp 1097–1105
40. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539
41. Li Deng DY (2014) Deep learning: methods and applications. NOW Publishers, Boston
42. Liu F, Shen C, Lin G (2015) Deep convolutional neural fields for depth estimation from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5162–5170
43. Liu H, Wu Y, Sun F, Fang B, Guo D (2018) Weakly paired multimodal fusion for object recognition. IEEE Trans Autom Sci Eng 15(2):784–795. https://doi.org/10.1109/TASE.2017.2692271
44. Logan I, Robert L, Humeau S, Singh S (2017) Multimodal attribute extraction. arXiv preprint arXiv:171111118
45. Lu X, Wu F, Li X, Zhang Y, Lu W, Wang D, Zhuang Y (2014) Learning multimodal neural network with ranking examples. In: Proceedings of the 22nd ACM international conference on multimedia. ACM, pp 985–988
46. Madjarov G, Kocev D, Gjorgjevikj D, Džeroski S (2012) An extensive experimental comparison of methods for multi-label learning. Pattern Recognit 45(9):3084–3104. https://doi.org/10.1016/j.patcog.2012.03.004
47. Makita E, Lenskiy A (2016) A movie genre prediction based on Multivariate Bernoulli model and genre correlations. arXiv preprint arXiv:160408608 (May), arxiv:1604.08608
48. Makita E, Lenskiy A (2016) A multinomial probabilistic model for movie genre predictions. arXiv preprint arXiv:160307849, http://arxiv.org/abs/1603.07849
49. Mandal D, Biswas S (2016) Generalized coupled dictionary learning approach with applications to cross-modal matching. IEEE Trans Image Process 25(8):3826–3837
50. Mao J, Xu W, Yang Y, Wang J, Yuille AL (2014) Explain images with multimodal recurrent neural networks. arXiv preprint arXiv:14101090
51. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781
52. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds) Advances in neural information processing systems. Curran Associates Inc, New york, pp 3111–3119
53. Ngiam J, Khosla A, Kim M (2011) Multimodal deep learning. In: Proceedings of the 28th international conference on machine learning (ICML-11), pp 689–696. http://ai.stanford.edu/~ang/papers/icml11-MultimodalDeepLearning.pdf. Accessed June 7 2018
54. Norouzi M, Mikolov T, Bengio S, Singer Y, Shlens J, Frome A, Corrado GS, Dean J (2014) Zero-shot learning by convex combination of semantic embeddings. CoRR abs/1312.5, arxiv:1312.5650
55. Pei D, Liu H, Liu Y, Sun F (2013) Unsupervised multimodal feature learning for semantic image segmentation. In: The 2013 international joint conference on neural networks (IJCNN). IEEE, pp 1–6. https://doi.org/10.1109/IJCNN.2013.6706748
56. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556
57. Socher R, Ganjoo M, Manning CD, Ng A (2013) Zero-shot learning through cross-modal transfer. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds) Advances in neural information processing systems, vol 26. Curran Associates Inc, Hook, pp 935–943
58. Socher R, Karpathy A, Le QV, Manning CD, Ng AY (2014) Grounded compositional semantics for finding and describing images with sentences. Trans Assoc Comput Linguist (TACL) 2:207–218
59. Srivastava N, Salakhutdinov R (2012) Multimodal learning with deep Boltzmann machines. In: Pereira F, Burges C, Bottou L, Weinberger K (eds) Advances in neural information processing systems, vol 25. Curran Associates Inc, Hook, pp 2222–2230
60. Srivastava RK, Greff K, Schmidhuber J (2015) Highway networks. arXiv preprint arXiv:150500387
61. Suk HI, Shen D (2013) Deep learning-based feature representation for AD/MCI classification. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol 8150. LNCS, pp 583–590. https://doi.org/10.1007/978-3-642-40763-5_72
62. Treml M, Arjona-Medina J, Unterthiner T, Durgesh R, Friedmann F, Schuberth P, Mayr A, Heusel M, Hofmarcher M, Widrich M et al (2016) Speeding up semantic segmentation for autonomous driving. NIPSW 1(7):8
63. Tu J, Wu Z, Dai Q, Jiang YG, Xue X (2014) Challenge Huawei challenge: fusing multimodal features with deep neural networks for mobile video annotation. In: 2014 IEEE international conference on multimedia and expo workshops (ICMEW), pp 1–6. https://doi.org/10.1109/ICMEW.2014.6890609
64. Valada A, Dhall A, Burgard W (2016) Convoluted mixture of deep experts for robust semantic segmentation. In: IEEE/RSJ international conference on intelligent robots and systems (IROS) workshop, state estimation and terrain perception for all terrain mobile robots
65. Valada A, Oliveira G, Brox T, Burgard W (2016) Deep multispectral semantic scene understanding of forested environments

using multimodal fusion. In: The 2016 international symposium on experimental robotics (ISER 2016), Tokyo, Japan. http://ais.informatik.uni-freiburg.de/publications/papers/valada16iser.pdf. Accessed June 7 2018
66. Van Merriënboer B, Bahdanau D, Dumoulin V, Serdyuk D, Warde-Farley D, Chorowski J, Bengio Y (2015) Blocks and fuel: frameworks for deep learning. arXiv preprint arXiv:150600619
67. Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: a neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3156–3164
68. Wei Q (2015) Bayesian fusion of multi-band images: a powerful tool for super-resolution. Ph.D. thesis, Institut National Polytechnique de Toulouse (INPT)
69. Wei Q, Dobigeon N, Tourneret JY (2015) Bayesian fusion of multi-band images. IEEE J Sel Top Signal Process 9(6):1117–1127
70. Wu P, Hoi SC, Xia H, Zhao P, Wang D, Miao C (2013) Online multimodal deep similarity learning with application to image retrieval. In: Proceedings of the 21st ACM international conference on multimedia—MM '13. ACM Press, New York, pp 153–162. https://doi.org/10.1145/2502081.2502112
71. Wu Q, Teney D, Wang P, Shen C, Dick A, van den Hengel A (2017) Visual question answering: a survey of methods and datasets. Comput Vis Image Underst 163:21–40
72. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel RS, Bengio Y (2015) Show, attend and tell: neural image caption generation with visual attention 2(3):5. arXiv preprint arXiv:150203044
73. Yan R, Zhao D (2018) Smarter response with proactive suggestion: a new generative neural conversation paradigm. In: IJCAI, pp 4525–4531
74. Yao L, Zhang Y, Feng Y, Zhao D, Yan R (2017) Towards implicit content-introducing for generative short-text conversation systems. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 2190–2199
75. Ye F, Pu J, Wang J, Li Y, Zha H (2017) Glioma grading based on 3d multimodal convolutional neural network and privileged learning. In: 2017 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE, pp 759–763
76. Yuksel SE, Wilson JN, Gader PD (2012) Twenty years of mixture of experts. IEEE Trans Neural Netw Learn Syst 23(8):1177–1193
77. Zhao J, Xie X, Xu X, Sun S (2017) Multi-view learning overview: recent progress and new challenges. Inf Fusion 38:43–54
78. Zheng Y, Zhang YJ, Larochelle H (2014) Topic modeling of multimodal data: an autoregressive approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1370–1377